Tutorial on Verification of Neural Networks

Stefanie Mühlberger

24.07.2020

Outline

- Preliminaries
- Overview of Approaches
  - MIP encoding
  - Reluplex
  - Interval Analysis
  - DeepPoly

Preliminaries

Neural Network

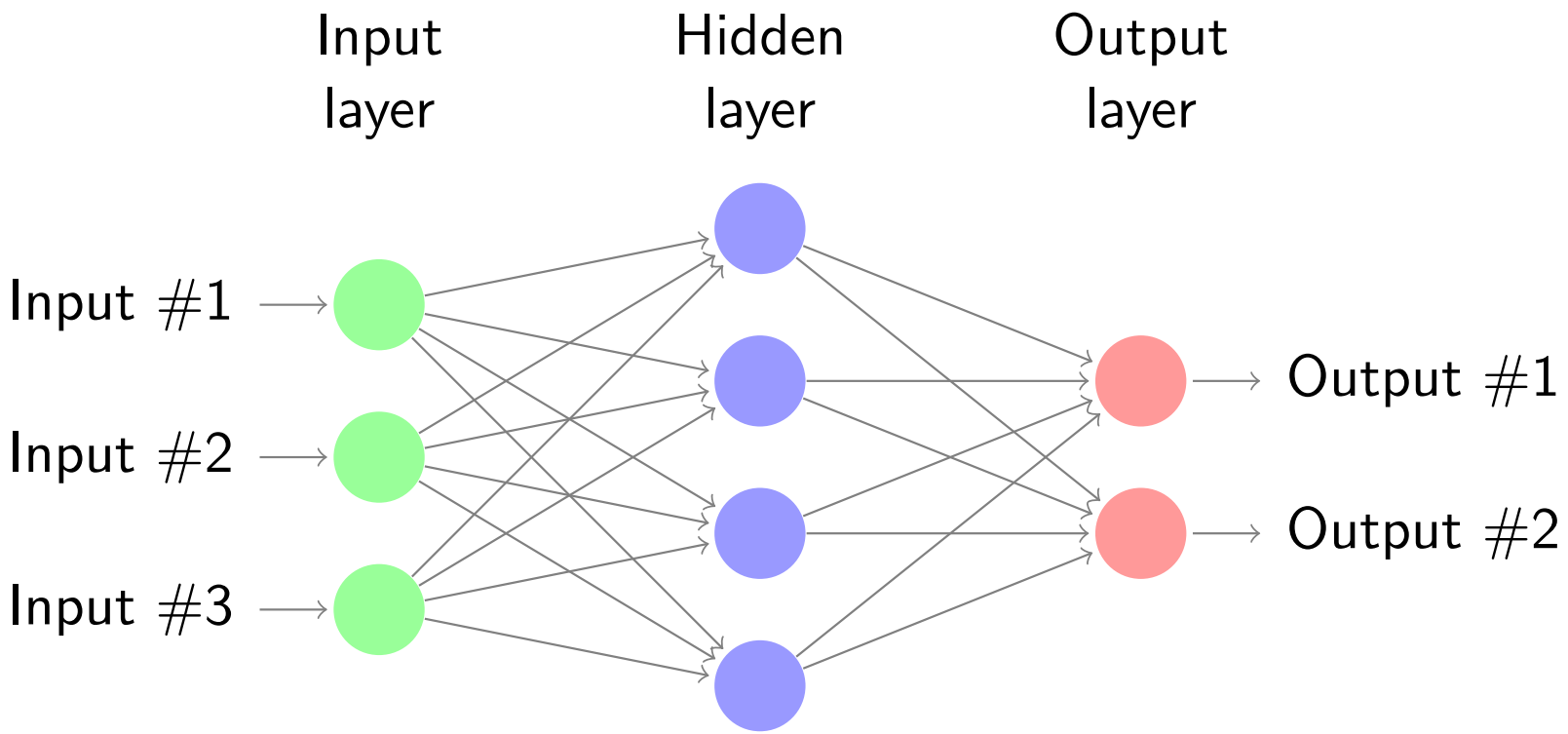


Figure: Neural Network

Preliminaries

Neuron

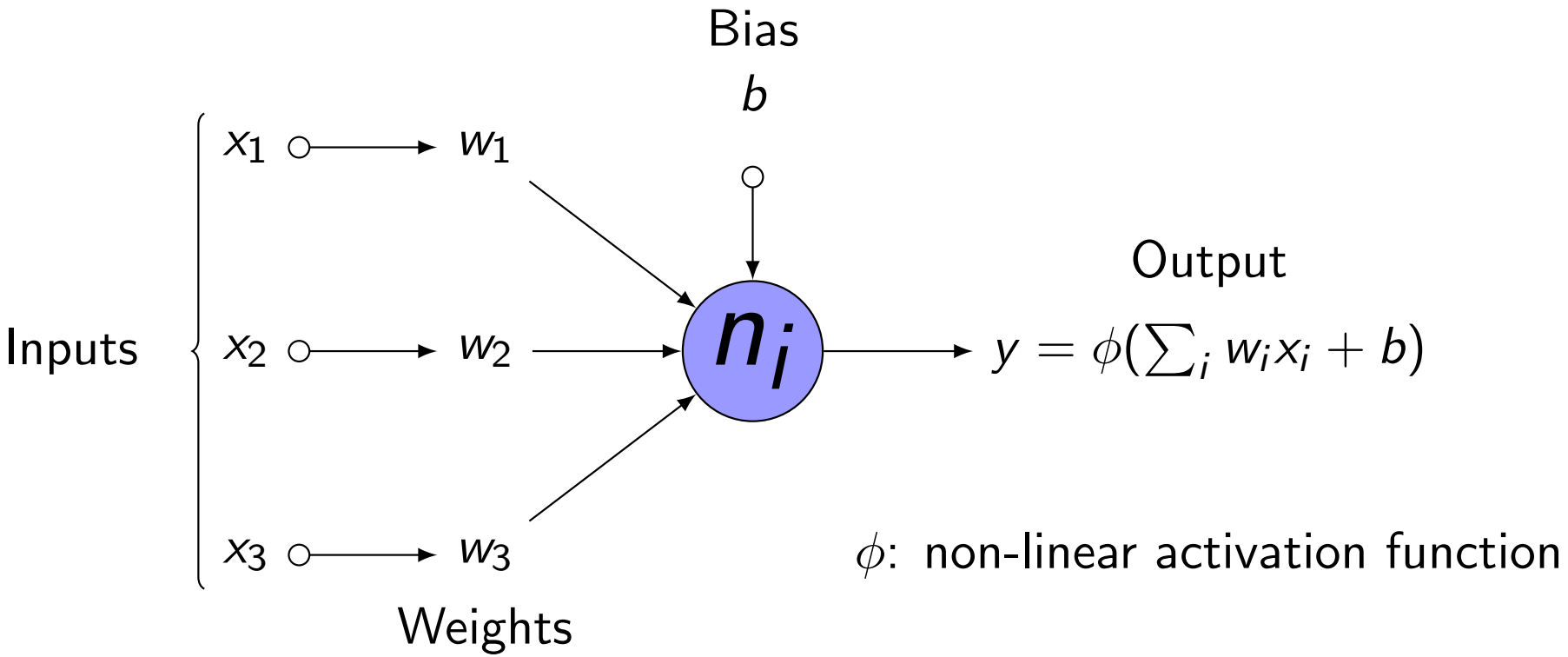


Figure: One Neuron n_i

Verification of Neural Networks

Problem

The behavior of NNs is hard to track.

We want to...

- ... evaluate its performance
- ... identify 'bad' behavior
- ... prove non-existance of 'bad' behavior

Verification of Neural Networks

Problem

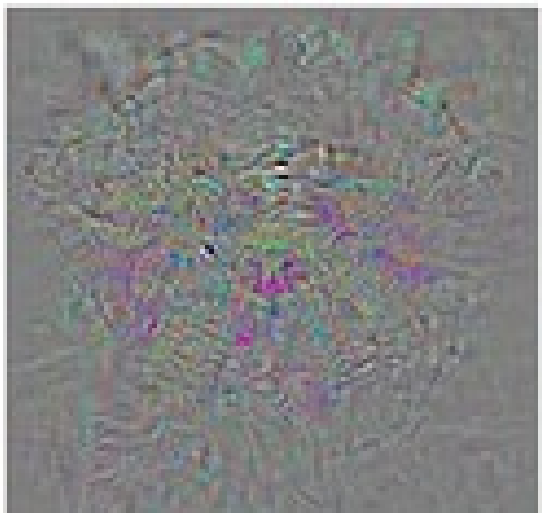
The behavior of NNs is hard to track.

We want to...

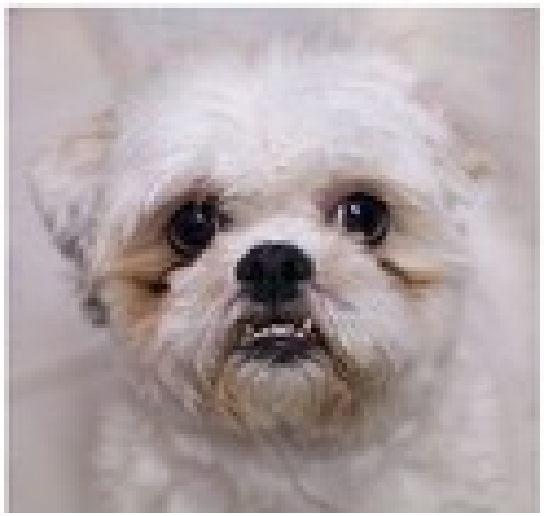
- ... evaluate its performance $\Rightarrow$ accuracy, error, inference time, ...
- ... identify 'bad' behavior $\Rightarrow$ attacks of NNs, ...
- ... prove non-existance of 'bad' behavior $\Rightarrow$ **verification of safety guarantees**

My Neural Network has been tricked!

dog



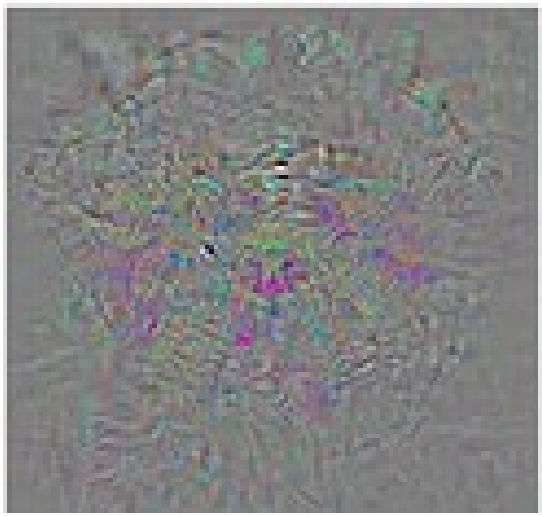
distortion



[9]

My Neural Network has been tricked!

dog



distortion

ostrich

Properties

Local Robustness

Given input x and a small distortion ϵ, the NN f should still predict something similar as x (or even the same).
$\forall \delta \; \exists \epsilon \; \forall x \in D \; \forall y \in B_\epsilon(x) \; : \; |f(x) - f(y)| \leq \delta$

General Input/Output-Constraint

Given input x and output $y = f(x)$ of some NN f, it shall fulfill some linear constraint $c(x, y)$.

Examples

- output larger than zero
- if the input is larger than zero, the output should also be larger
- the output should be greater than the input

Verification Approaches

Complete

\+ counterexample

\- hardly scalable

Examples

- Planet [3]
- Reluplex [5]
- MIP encoding [2][11]

Incomplete

\+ faster - no counterexample

\- overapproximation

Examples

- Reachability Analysis [7]
- Interval Analysis [12]
- Abstract Domain Analysis [4][8]

MIP

Rewrite the Neural Network f in linear constraints [2]

Optimization

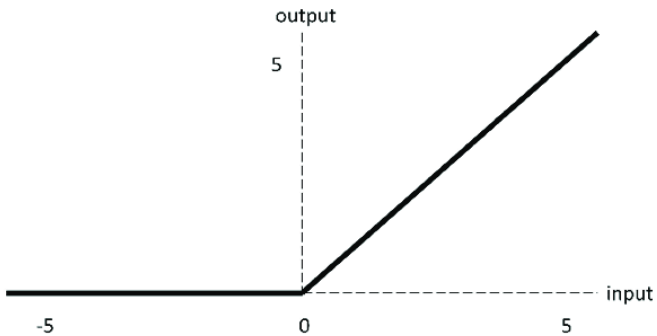
$\min_{x'} d(x', x)$
subject to $f(x') \neq f(x)$
and $x' \in X_{valid}$

X_{valid}

- For all layers $l = 1, ..., L$ with n_l neurons
- $im_i^l = \sum_{i=1}^{n_{l-1}} w_{ij} \cdot out_i^{l-1}$
- $out_i^l = \phi(im_i^l)$ where ϕ is the activation function

Non-linear Activation Function

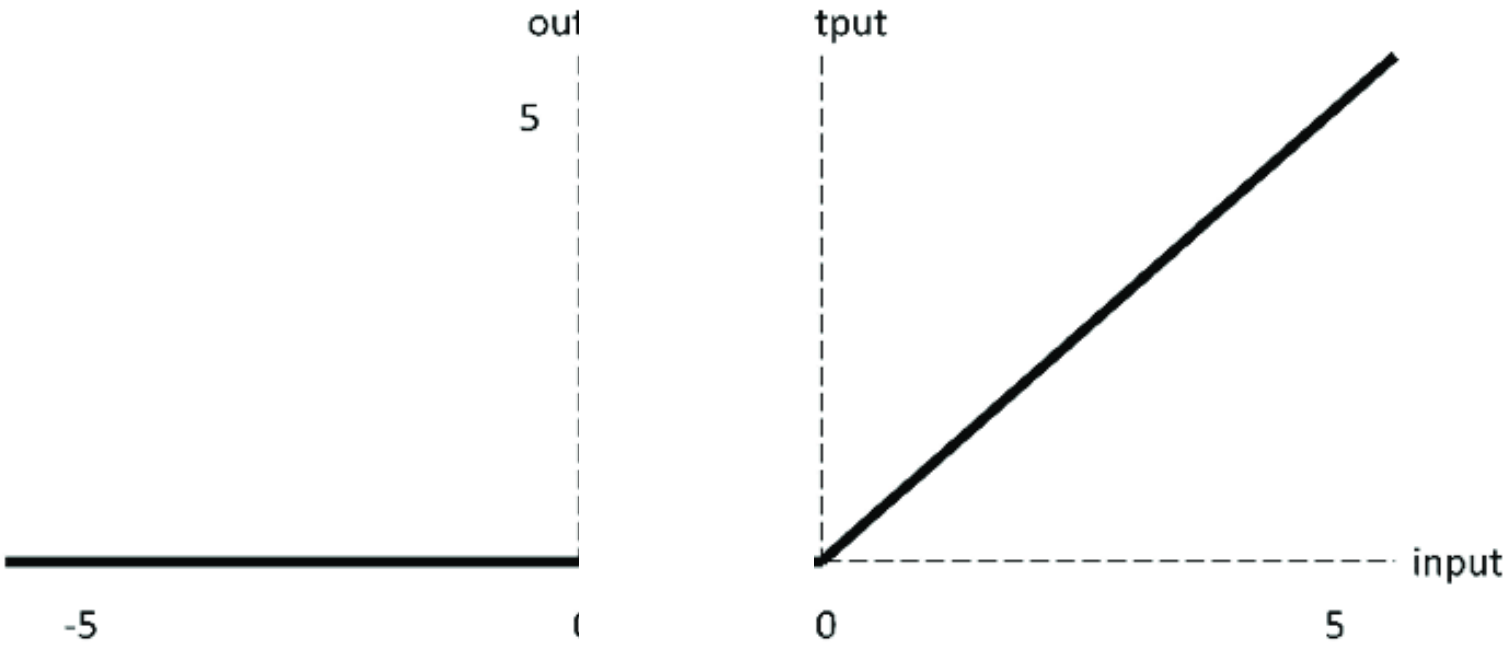
$$\text{ReLU}(x) = \max(x, 0)$$



Isn't that almost linear?

Non-linear Activation Function

ReLU



Let's just split the ReLU!
Branch the verification algorithm

MIP

Problems

- Branching $\Rightarrow$ Exponential Blow-up
- Only ReLU as activation-function

MIP

Smarter Idea [2]:

ReLU

Big-M encoding to replace disjunctions:
$y = \max(x, 0)$ is equivalent to

$$
\begin{aligned}
y &\geq 0 \\
y &\geq x \\
x - bM &\leq 0 \\
x + (1 - b)M &\geq 0 \\
y &\leq x + (1 - b)M \\
y &\leq bM
\end{aligned}
$$

where b is a binary integer variable

MIP

Problems

- Hardly scalable
- Calculation of good M is crucial

Reluplex

Reluplex

Encoding of a Neural Network into a **SMT** instance and solving it with a modified version of the **Simplex algorithm**, by adding an additional ReLU-constraint.

Reluplex

SMT

Satisfiability Modulo Theories

Check satisfiablitiy of logic atoms which are built on a theory (e.g. linear arithmetic).

Simplex

Algorithm to solve a set of linear constraints

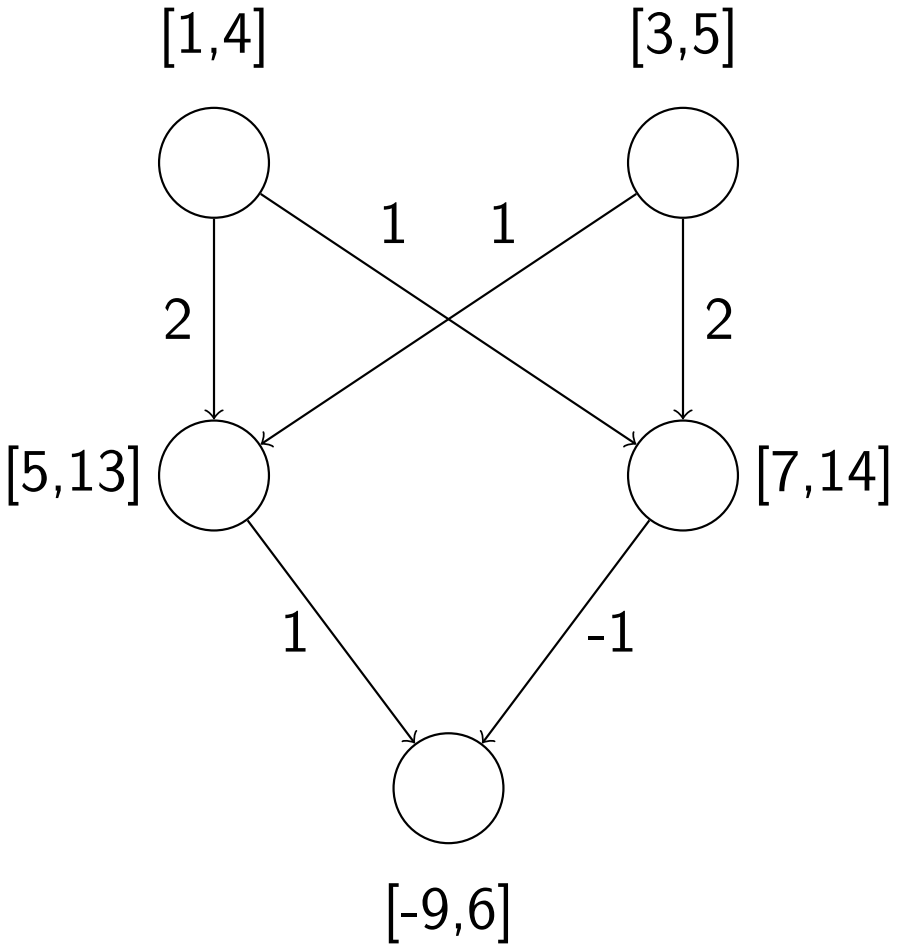
Reluplex

Difference to MIP

- do not split on all ReLU-nodes
- only split on 'problematic' nodes

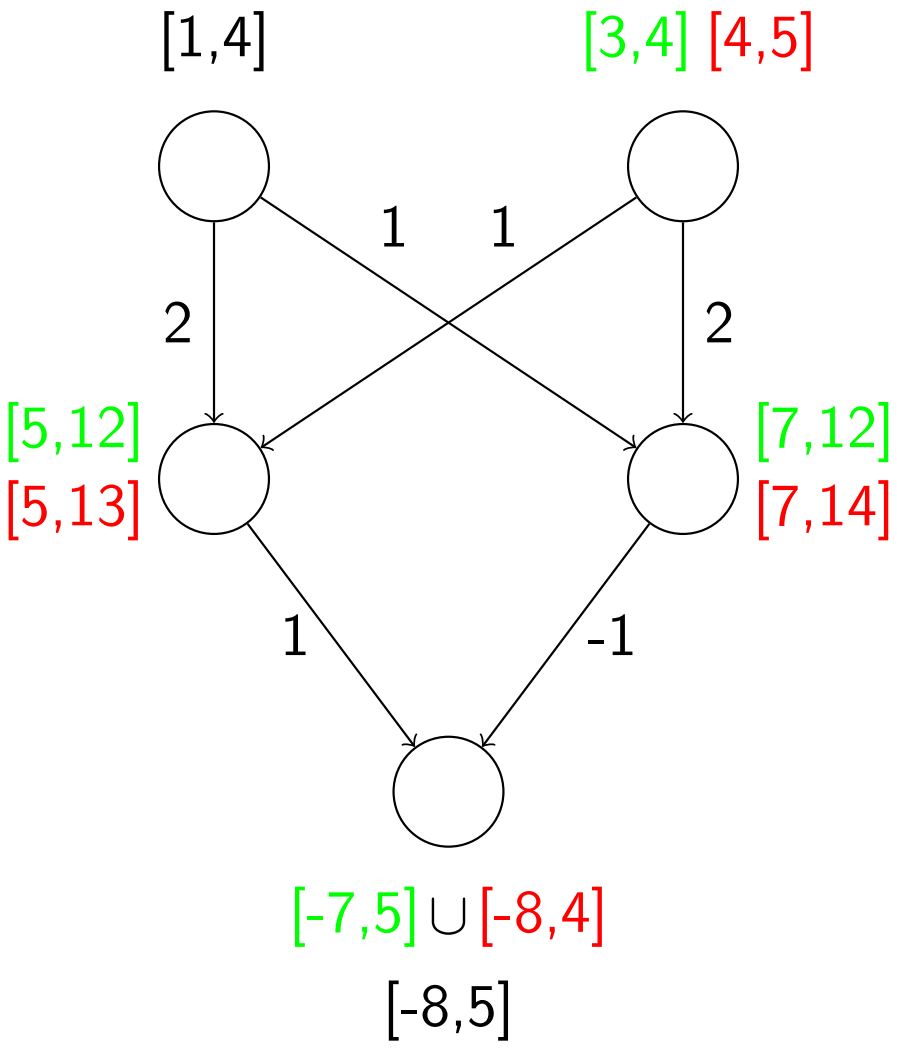
Interval Propagation

Naive idea: propagate the input interval through the network [3]



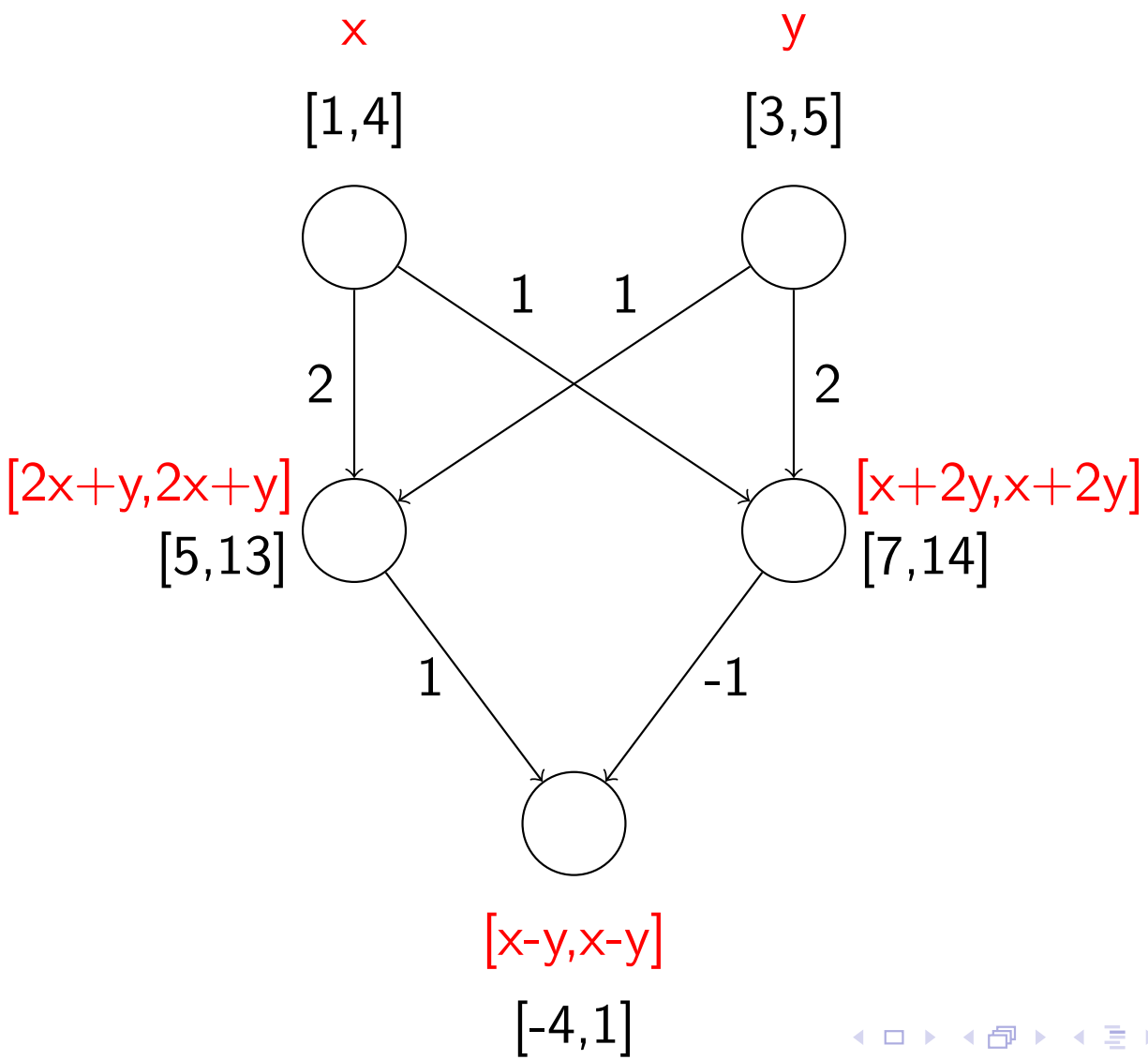
Interval Propagation

Smarter idea: Bisection of input features [12]



Interval Propagation

Smarter idea: propagate the symbolic interval through the network [12]



DeepPoly

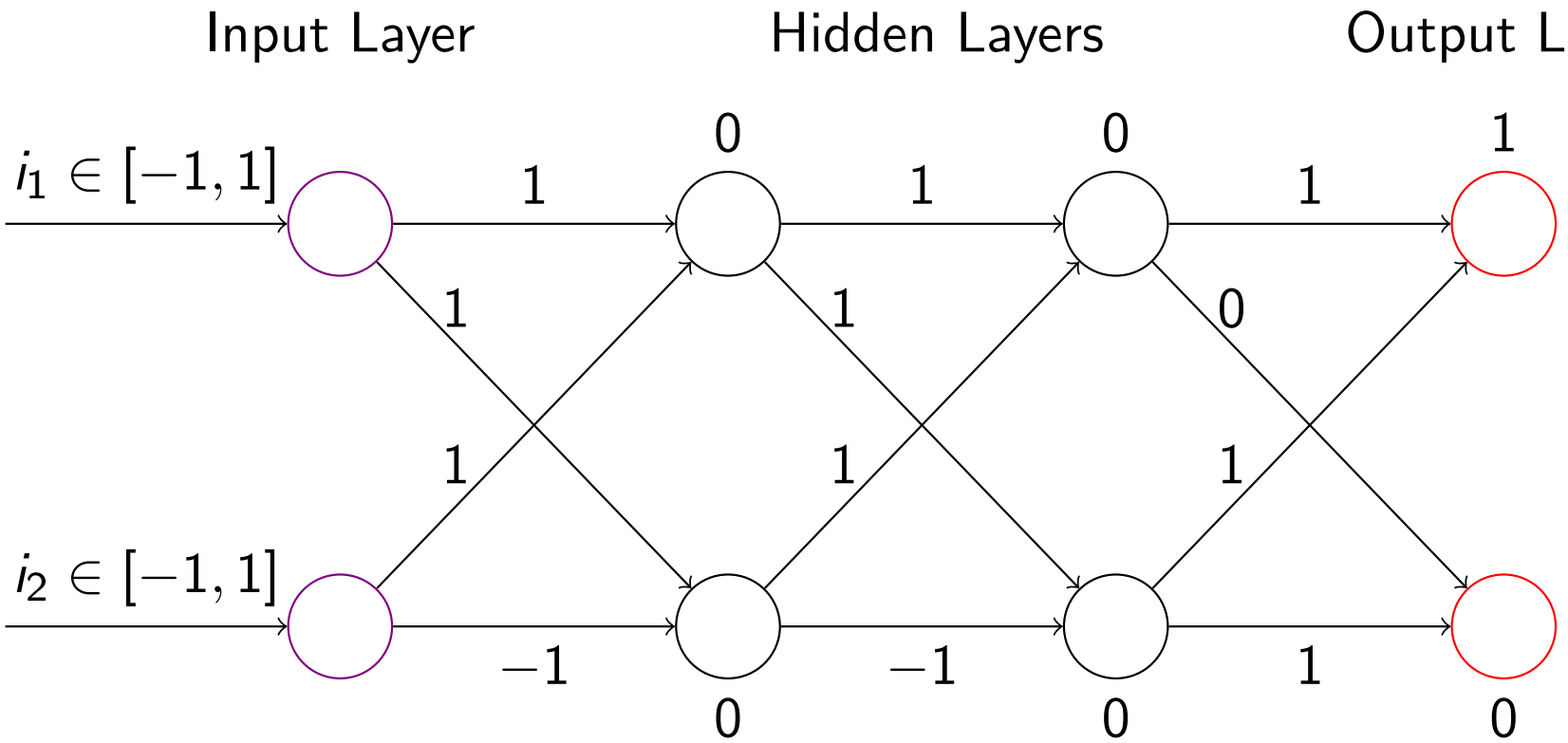
Even smarter: Don't use just intervals, but a combination of symbolic and real intervals, namely an abstract domain (polyhedra).

$$a = \langle a^{\leq}, a^{\geq}, l, u \rangle$$

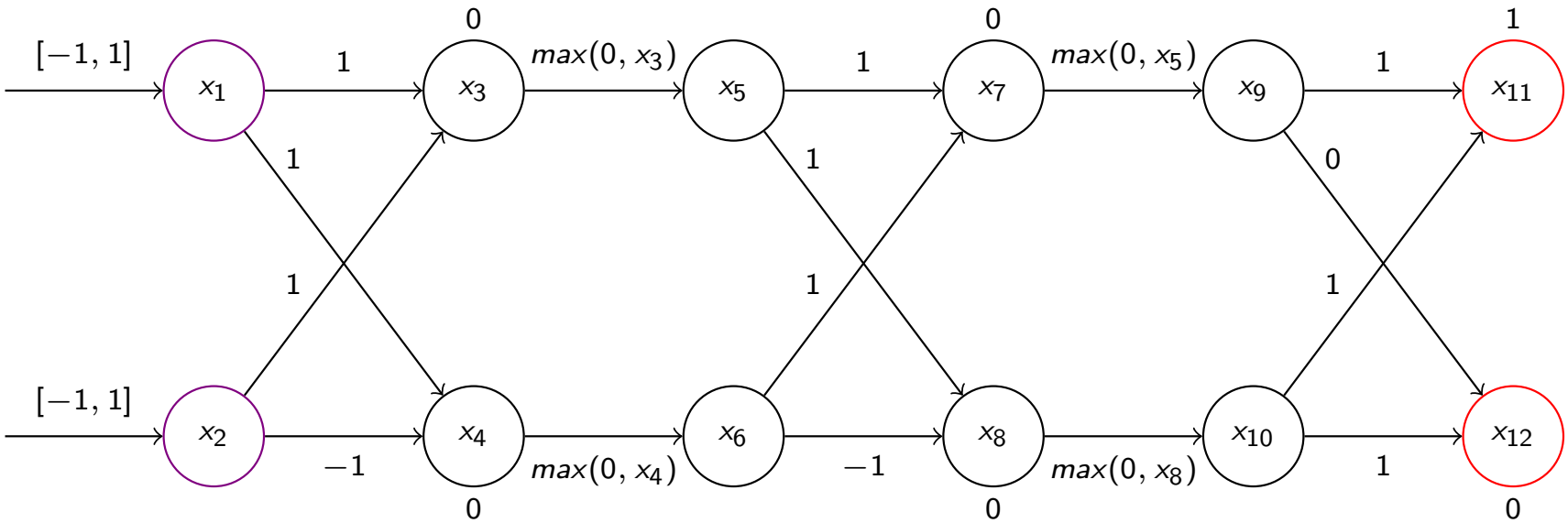
where $a^{\leq}, a^{\geq}$ are constraints on n variables $x_1, ..., x_n$.
More precisely

$$a_i^{\leq}, a_i^{\geq} \in \left\{ v + \sum_{j=1}^{i-1} w_j \cdot x_j \mid v \in \mathbb{R} \cup \{-\infty, +\infty\}, w \in \mathbb{R}^{i-1} \right\} \text{ for } i \in \{1, ..., n\}$$

DeepPoly - Example

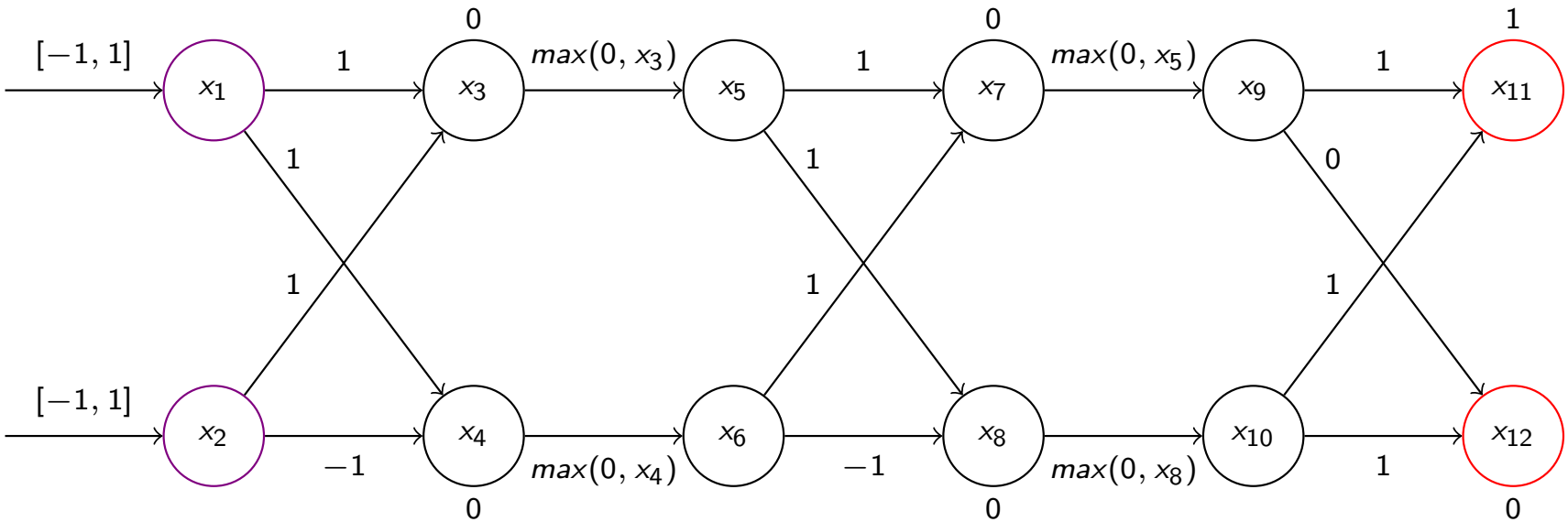


DeepPoly - Example



DeepPoly - Example

$(x_1 \geq -1,$
$x_1 \leq 1,$
$l_1 = -1,$
$u_1 = 1)$

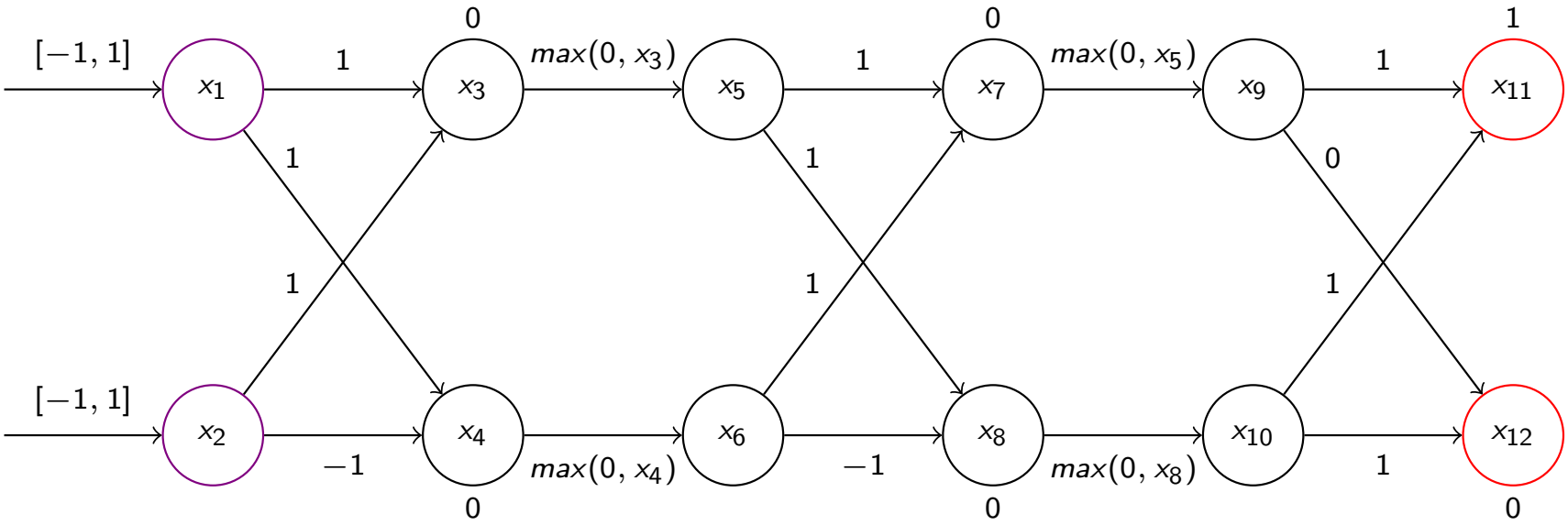


$(x_2 \geq -1,$
$x_2 \leq 1,$
$l_2 = -1,$
$u_2 = 1)$

DeepPoly - Example

$(x_1 \geq -1,$
$x_1 \leq 1,$
$l_1 = -1,$
$u_1 = 1)$

$(x_3 \geq x_1 + x_2,$
$x_3 \leq x_1 + x_2,$
$l_3 = -2,$
$u_3 = 2)$

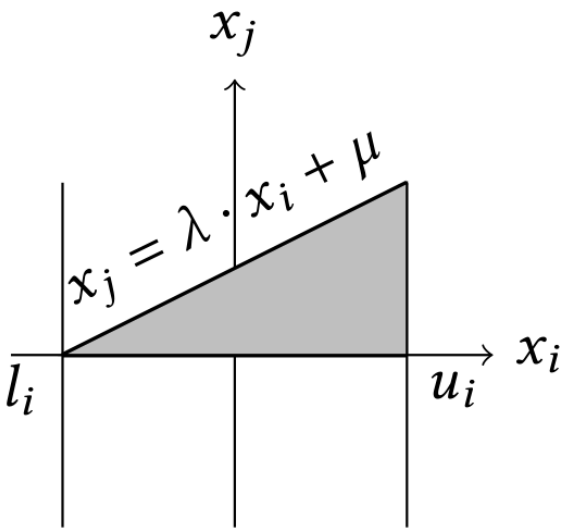


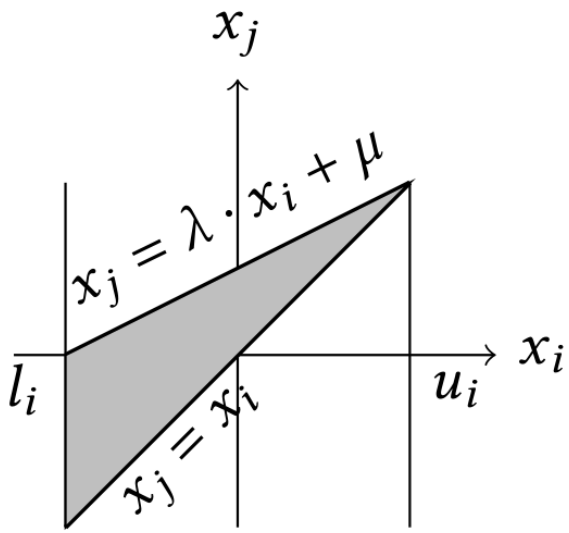
$(x_2 \geq -1,$
$x_2 \leq 1,$
$l_1 = -1,$
$u_1 = 1)$

$(x_4 \geq x_1 - x_2,$
$x_4 \leq x_1 - x_2,$
$l_3 = -2,$
$u_3 = 2)$

DeepPoly - Example

Overapproximation of the ReLU





where $\lambda = \frac{u_i}{u_i - l_i}$ and $\mu = \frac{-l_i u_i}{u_i - l_i}$

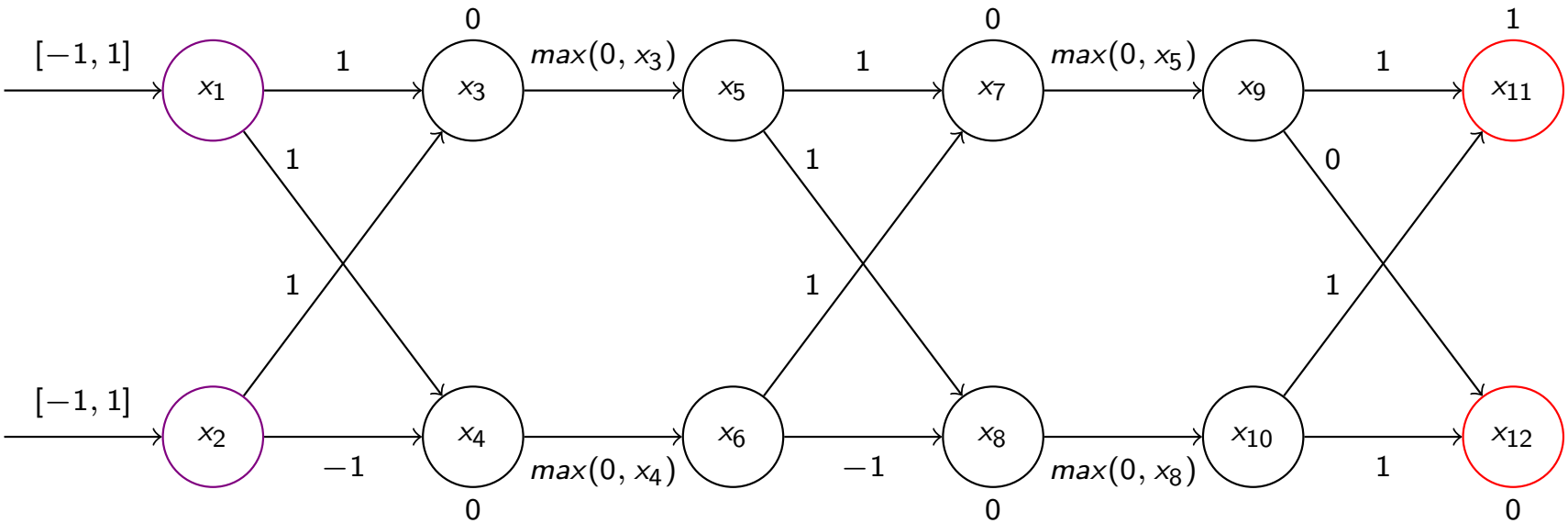
Choose whichever has the smaller area.

DeepPoly - Example

$(x_1 \geq -1,$
$x_1 \leq 1,$
$l_1 = -1,$
$u_1 = 1)$

$(x_3 \geq x_1 + x_2,$
$x_3 \leq x_1 + x_2,$
$l_3 = -2,$
$u_3 = 2)$

$(x_5 \geq 0,$
$x_5 \leq 0.5 \cdot x_3 + 1,$
$l_5 = 0,$
$u_5 = 2)$



$(x_2 \geq -1,$
$x_2 \leq 1,$
$l_1 = -1,$
$u_1 = 1)$

$(x_4 \geq x_1 - x_2,$
$x_4 \leq x_1 - x_2,$
$l_3 = -2,$
$u_3 = 2)$

$(x_6 \geq 0,$
$x_6 \leq 0.5 \cdot x_4 + 1,$
$l_6 = 0,$
$u_6 = 2)$

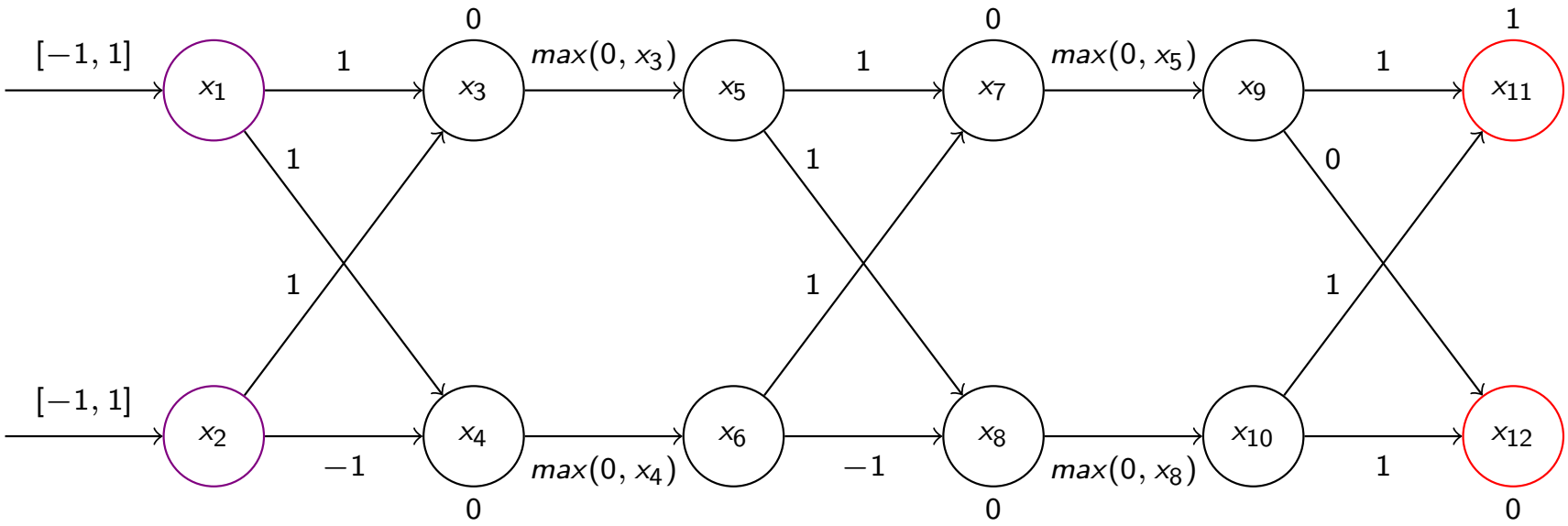
DeepPoly - Example

$(x_1 \geq -1,$
$x_1 \leq 1,$
$l_1 = -1,$
$u_1 = 1)$

$(x_3 \geq x_1 + x_2,$
$x_3 \leq x_1 + x_2,$
$l_3 = -2,$
$u_3 = 2)$

$(x_5 \geq 0,$
$x_5 \leq 0.5 \cdot x_3 + 1,$
$l_5 = 0,$
$u_5 = 2)$

$(x_7 \geq x_5 + x_6,$
$x_7 \leq x_5 + x_6,$
$l_7 = 0,$
$u_7 = 3)$



$(x_2 \geq -1,$
$x_2 \leq 1,$
$l_1 = -1,$
$u_1 = 1)$

$(x_4 \geq x_1 - x_2,$
$x_4 \leq x_1 - x_2,$
$l_3 = -2,$
$u_3 = 2)$

$(x_6 \geq 0,$
$x_6 \leq 0.5 \cdot x_4 + 1,$
$l_6 = 0,$
$u_6 = 2)$

$(x_8 \geq x_5 - x_6,$
$x_8 \leq x_5 - x_6,$
$l_8 = -2,$
$u_8 = 2)$

DeepPoly - Example

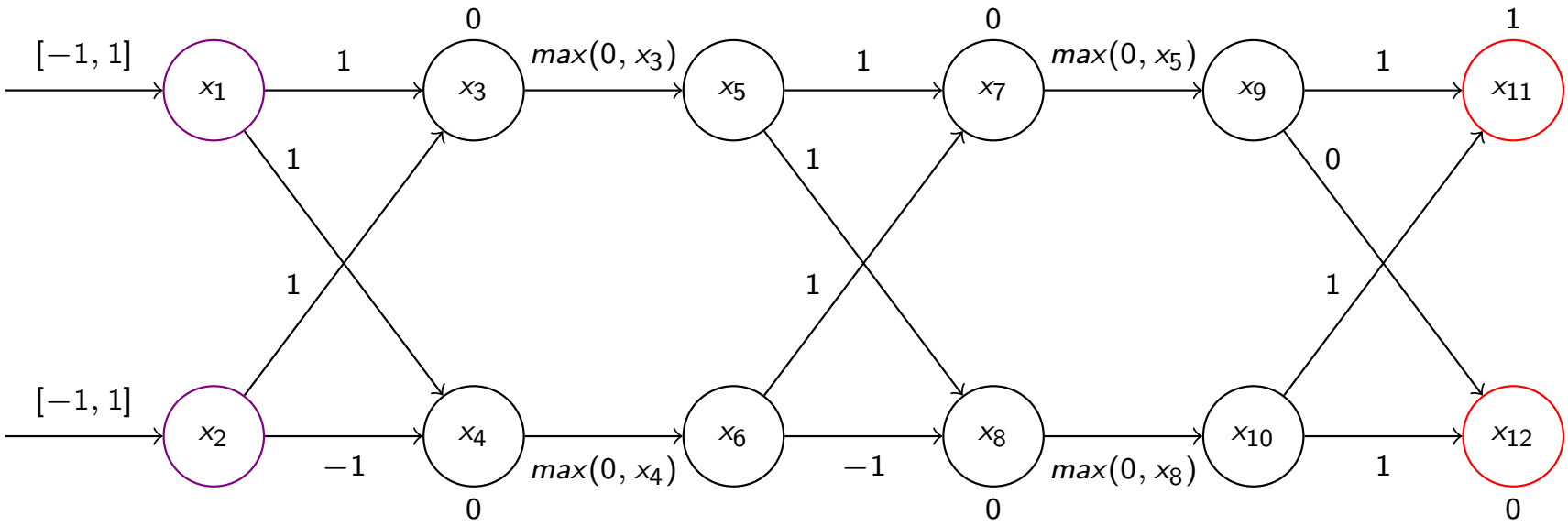
$(x_1 \geq -1,$
$x_1 \leq 1,$
$l_1 = -1,$
$u_1 = 1)$

$(x_3 \geq x_1 + x_2,$
$x_3 \leq x_1 + x_2,$
$l_3 = -2,$
$u_3 = 2)$

$(x_5 \geq 0,$
$x_5 \leq 0.5 \cdot x_3 + 1,$
$l_5 = 0,$
$u_5 = 2)$

$(x_7 \geq x_5 + x_6,$
$x_7 \leq x_5 + x_6,$
$l_7 = 0,$
$u_7 = 3)$

$(x_9 \geq x_7,$
$x_9 \leq x_7,$
$l_9 = 0,$
$u_9 = 3)$



$(x_2 \geq -1,$
$x_2 \leq 1,$
$l_1 = -1,$
$u_1 = 1)$

$(x_4 \geq x_1 - x_2,$
$x_4 \leq x_1 - x_2,$
$l_3 = -2,$
$u_3 = 2)$

$(x_6 \geq 0,$
$x_6 \leq 0.5 \cdot x_4 + 1,$
$l_6 = 0,$
$u_6 = 2)$

$(x_8 \geq x_5 - x_6,$
$x_8 \leq x_5 - x_6,$
$l_8 = -2,$
$u_8 = 2)$

$(x_{10} \geq 0,$
$x_{10} \leq 0.5 \cdot x_8 + 1,$
$l_{10} = 0,$
$u_{10} = 2)$

DeepPoly - Example

$(x_{11} \geq x_9 + x_{10} + 1,$
$x_{11} \leq x_9 + x_{10},$
$l_{11} = 1,$
$u_{11} = 5.5)$

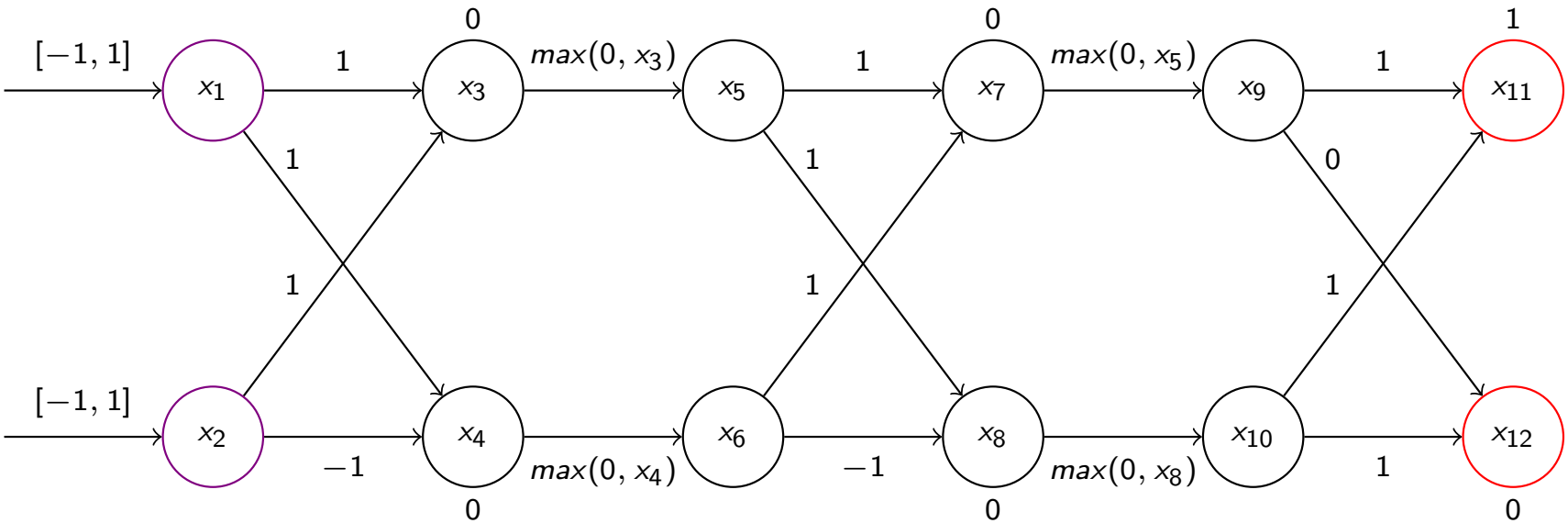
$(x_1 \geq -1,$
$x_1 \leq 1,$
$l_1 = -1,$
$u_1 = 1)$

$(x_3 \geq x_1 + x_2,$
$x_3 \leq x_1 + x_2,$
$l_3 = -2,$
$u_3 = 2)$

$(x_5 \geq 0,$
$x_5 \leq 0.5 \cdot x_3 + 1,$
$l_5 = 0,$
$u_5 = 2)$

$(x_7 \geq x_5 + x_6,$
$x_7 \leq x_5 + x_6,$
$l_7 = 0,$
$u_7 = 3)$

$(x_9 \geq x_7,$
$x_9 \leq x_7,$
$l_9 = 0,$
$u_9 = 3)$



$(x_2 \geq -1,$
$x_2 \leq 1,$
$l_1 = -1,$
$u_1 = 1)$

$(x_4 \geq x_1 - x_2,$
$x_4 \leq x_1 - x_2,$
$l_3 = -2,$
$u_3 = 2)$

$(x_6 \geq 0,$
$x_6 \leq 0.5 \cdot x_4 + 1,$
$l_6 = 0,$
$u_6 = 2)$

$(x_8 \geq x_5 - x_6,$
$x_8 \leq x_5 - x_6,$
$l_8 = -2,$
$u_8 = 2)$

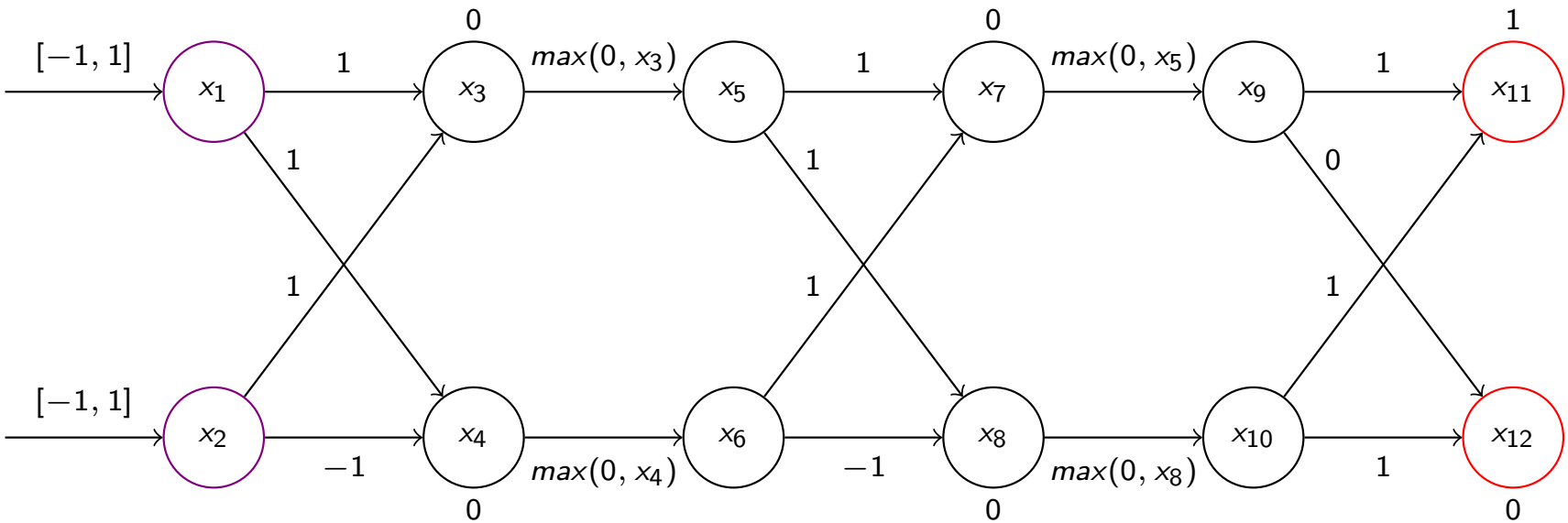
$(x_{10} \geq 0,$
$x_{10} \leq 0.5 \cdot x_8 + 1,$
$l_{10} = 0,$
$u_{10} = 2)$

$(x_{12} \geq x_{10},$
$x_{12} \leq x_{10},$
$l_{12} = 0,$
$u_{12} = 2)$

DeepPoly - Example

If we want to check, whether $x_{11} < x_{12}$ or $x_{11} > x_{12}$, we can't conclude anything ☹.

$(x_1 \geq -1, x_1 \leq 1, l_1 = -1, u_1 = 1)$ $(x_3 \geq x_1 + x_2, x_3 \leq x_1 + x_2, l_3 = -2, u_3 = 2)$ $(x_5 \geq 0, x_5 \leq 0.5 \cdot x_3 + 1, l_5 = 0, u_5 = 2)$ $(x_7 \geq x_5 + x_6, x_7 \leq x_5 + x_6, l_7 = 0, u_7 = 3)$ $(x_9 \geq x_7, x_9 \leq x_7, l_9 = 0, u_9 = 3)$ $(x_{11} \geq x_9 + x_{10} + 1, x_{11} \leq x_9 + x_{10}, l_{11} = 1, u_{11} = 5.5)$



$(x_2 \geq -1, x_2 \leq 1, l_1 = -1, u_1 = 1)$ $(x_4 \geq x_1 - x_2, x_4 \leq x_1 - x_2, l_3 = -2, u_3 = 2)$ $(x_6 \geq 0, x_6 \leq 0.5 \cdot x_4 + 1, l_6 = 0, u_6 = 2)$ $(x_8 \geq x_5 - x_6, x_8 \leq x_5 - x_6, l_8 = -2, u_8 = 2)$ $(x_{10} \geq 0, x_{10} \leq 0.5 \cdot x_8 + 1, l_{10} = 0, u_{10} = 2)$ $(x_{12} \geq x_{10}, x_{12} \leq x_{10}, l_{12} = 0, u_{12} = 2)$

DeepPoly - Example

Set $x_{13} = x_{11} - x_{12}$.
Backtrack $x_{13} \geq x_{11} - x_{12}$

$$\begin{aligned} x_{13} &\geq x_{11} - x_{12} \\ &\geq x_9 + x_{10} + 1 - x_{10} \\ &\geq x_7 + 1 \\ &\geq x_5 + x_6 + 1 \\ &\geq 1 \end{aligned}$$

Thus, $x_{11} > x_{12}$ ☺.

Conclusion

Ongoing Work

Still new approaches and ideas (e.g. ImageStars[10] from CAV2020)

Take Away

For smaller networks: go for a complete approach
For large networks: use abstraction techniques [1][6]and/or use incomplete approaches

Thanks to machine-learning algorithms,
the robot apocalypse was short-lived.

References I

Pranav Ashok et al. "DeepAbstract: Neural Network Abstraction for Accelerating Verification". In: *to be in ATVA* (2020). arXiv: 2006.13735. URL: https://arxiv.org/abs/2006.13735.

Chih-Hong Cheng, Georg Nührenberg, and Harald Ruess. "Maximum Resilience of Artificial Neural Networks". In: *Automated Technology for Verification and Analysis - 15th International Symposium, ATVA 2017, Pune, India, October 3-6, 2017, Proceedings*. Ed. by Deepak D'Souza and K. Narayan Kumar. Vol. 10482. Lecture Notes in Computer Science. Springer, 2017, pp. 251–268. DOI: 10.1007/978-3-319-68167-2_18. URL: https://doi.org/10.1007/978-3-319-68167-2%5C_18.

Rüdiger Ehlers. "Formal Verification of Piece-Wise Linear Feed-Forward Neural Networks". In: *Automated Technology for Verification and Analysis - 15th International Symposium, ATVA 2017, Pune, India, October 3-6, 2017, Proceedings*. Ed. by Deepak D'Souza and K. Narayan Kumar. Vol. 10482. Lecture Notes in Computer Science. Springer, 2017, pp. 269–286. DOI: 10.1007/978-3-319-68167-2_19. URL: https://doi.org/10.1007/978-3-319-68167-2%5C_19.

Timon Gehr et al. "AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation". In: *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*. IEEE Computer Society, 2018, pp. 3–18. DOI: 10.1109/SP.2018.00058. URL: https://doi.org/10.1109/SP.2018.00058.

Guy Katz et al. "Reluplex: An efficient SMT solver for verifying deep neural networks". In: *International Conference on Computer Aided Verification*. Springer. 2017, pp. 97–117.

References II

Guy Katz et al. "The Marabou Framework for Verification and Analysis of Deep Neural Networks". In: *Computer Aided Verification - 31st International Conference, CAV 2019, New York City, NY, USA, July 15-18, 2019, Proceedings, Part I*. Ed. by Isil Dillig and Serdar Tasiran. Vol. 11561. Lecture Notes in Computer Science. Springer, 2019, pp. 443–452. DOI: 10.1007/978-3-030-25540-4_26. URL: https://doi.org/10.1007/978-3-030-25540-4%5C_26.

Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. "Reachability Analysis of Deep Neural Networks with Provable Guarantees". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*. Ed. by Jérôme Lang. ijcai.org, 2018, pp. 2651–2659. DOI: 10.24963/ijcai.2018/368. URL: https://doi.org/10.24963/ijcai.2018/368.

Gagandeep Singh et al. "An abstract domain for certifying neural networks". In: *Proc. ACM Program. Lang.* 3.POPL (2019), 41:1–41:30. DOI: 10.1145/3290354. URL: https://doi.org/10.1145/3290354.

Christian Szegedy et al. "Intriguing properties of neural networks". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2014. URL: http://arxiv.org/abs/1312.6199.

Hoang-Dung Tran et al. "Verification of Deep Convolutional Neural Networks Using ImageStars". In: *Computer Aided Verification - 32nd International Conference, CAV 2020, Los Angeles, CA, USA, July 21-24, 2020, Proceedings, Part I*. Ed. by Shuvendu K. Lahiri and Chao Wang. Vol. 12224. Lecture Notes in Computer Science. Springer, 2020, pp. 18–42. DOI: 10.1007/978-3-030-53288-8_2. URL: https://doi.org/10.1007/978-3-030-53288-8%5C_2.

References III

Vincent Tjeng, Kai Y. Xiao, and Russ Tedrake. "Evaluating Robustness of Neural Networks with Mixed Integer Programming". In: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL: https://openreview.net/forum?id=HyGIdiRqtm.

Shiqi Wang et al. "Efficient Formal Safety Analysis of Neural Networks". In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*. Ed. by Samy Bengio et al. 2018, pp. 6369–6379. URL: http://papers.nips.cc/paper/7873-efficient-formal-safety-analysis-of-neural-networks.